4

Overview of Artificial Intelligence in Industrial-Organizational Psychology

Emily D. Campion and Michael A. Campion

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Introduction

Even in the short time that Industrial-Organizational psychologists have adopted and applied artificial intelligence (AI) to improve human resources decision making (Campion et al., 2016), significant progress has been made that foreshadows a promising future. The goal of this chapter is to offer an overview of how AI—and more specifically machine learning (ML)—has been used to advance how we measure work-related psychological constructs and improve organizational decision-making. We aim to make three points from our review and experience using ML in practice and research.

First, evidence to date on the value of ML in I-O psychology is encouraging, especially in personnel selection. The gains of using ML over traditional methods are most impressive when we can use it to analyze text data, which has typically been too resource-intensive to score for most organizations with only human raters. Yet, more research is needed (Landers et al., 2023). In light of this, we offer an overview of how AI has been used in I-O psychology through a review of the I-O psychology literature (N = 38 publications) and provide a primer on AI methods.

Second, these methods have become more accessible with greater realized availability of repositories such as GitHub, but also the advent, popularization, and advancements of large language models (LLMs). LLMs are language-based machine learning models built on potentially billions of parameters that produce text in response to a human prompt. The accessibility of these methods offered by LLMs and the like may be one of the most significant forthcoming advancements and continues an early maxim proffered by Grace Hopper—computer programming pioneer—promoting that computer coding be developed using English language, or something similar (Tietjen, 2022).

Finally, looking forward, we argue that over-reliance on AI systems and explainability of some ML algorithms, such as those that use more advanced modeling (e.g., transformers and neural networks; also LLMs), are two of our most formidable challenges, especially when used in high-stakes decision-making (e.g., selection,

promotion, turnover). This should not prohibit our continued pursuit of understanding how these models operate but should instead guide our critical analysis of when to use and when not to use these methods.

Review of Literature on Artificial Intelligence in Industrial-Organizational Psychology

We conducted a journal-based review (Hiebl, 2023) of the literature that explicitly used AI in I-O psychology. We searched for terms such as "artificial intelligence," "machine learning," and "natural language processing" in the following journals (in alphabetical order): *Human Resource Management Journal*, *Human Resource Management Review*, *International Journal of Selection and Assessment*, *Journal of Applied Psychology*, *Journal of Business and Psychology*, *Journal of Management*, *Organizational Research Methods*, *Personnel Psychology*, and *Research in Personnel and Human Resources Management*. This search yielded 38 journal articles, which are annotated in Table 4.2.

A Very Brief Primer on Artificial Intelligence Methods

AI broadly refers to methods that simulate human intelligence. This includes simple tasks like differentially weighting variables according to their relationships with a criterion to classify or predict an outcome, but it can also include more complex tasks, such as generating new content, photos, or videos (e.g., LLMs). A visual representing the relationships among the common AI methods used in I-O psychology is presented in Figure 4.1. A short glossary of common terms is provided in Table 4.1. While AI is executed using a family of methods, the most common AI method in I-O psychology presently is ML. ML refers to "a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention" (SAS, 2019).

The majority of research on AI in I-O psychology has leveraged text data, rather than using only already-quantified data (e.g., assessment scores). I-O psychologists have crafted and validated close-ended assessments that yield quantitative scores such that those assessments can inform many organizational decision-making reliably. Thus, the use of AI for evaluating already-quantified data did not offer an opportunity to advance I-O psychology. However, consistently scoring text-based data has been a major challenge given the enormous resource cost of doing so in selection, performance management, career development, and the like, and AI has enabled the automated scoring of text sources (corpuses) such as job applications, resumes, essays, job descriptions, and the like. As such, this information had been inadvertently neglected until the more recent, broader accessibility of AI-based tools available to practitioners and researchers that have potential to develop, validate,

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

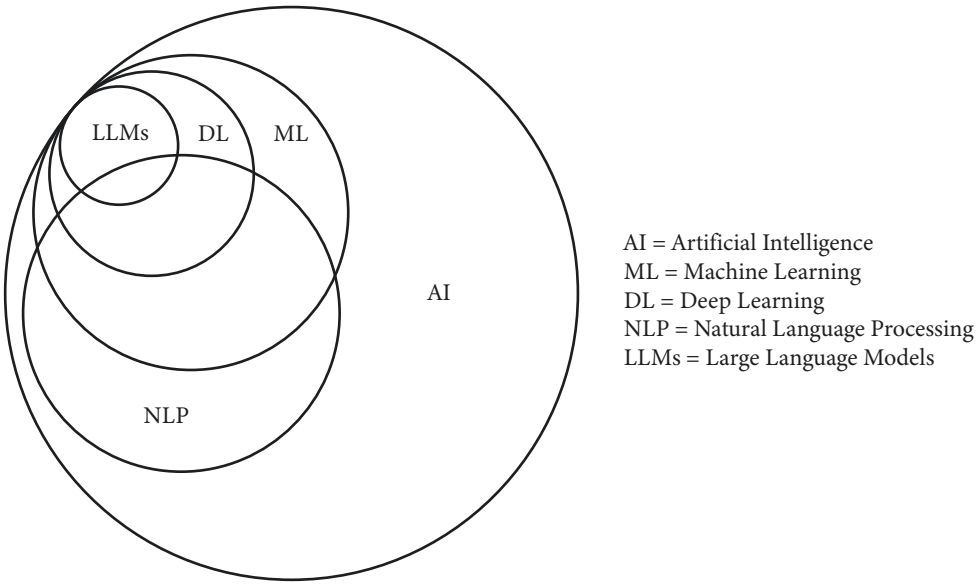


Figure 4.1 Venn Diagram of Relevant AI Methods in Industrial-Organizational Psychology
Source: Adapted from K. Silva (2021).

and operationalize AI-based scoring models. Because organizations collect enormous amounts of text data that have historically not been systematically integrated into organizational decision making, it may be that text analysis could become the primary analytic approach.

To analyze text data necessitates natural language processing (NLP), or algorithms specifically designed to enable machines to understand and generate human language. In NLP, language is preprocessed by first reducing corpi into n-grams. *N*-grams refer to single (e.g., "leader") or multi-terms (e.g., bi-gram; "strong leaders") that are extracted from a corpus. These are not directly entered into the model without first standardizing the language, which is typically done through stemming or lemmatizing. Stemming is a simpler standardization method where suffixes are removed from terms. For example, "lead," "leader," "leaders," and "leadership" would likely all become "lead." Lemmatizing retains terms in their dictionary form such that "lead," "leader," and "leadership" would stay the same, but "leaders" would become "leader." *N*-grams are then vectorized (transformed into an algorithm-readable matrix) in a number of ways. Bag-of-Words (BoW) is one of the earliest and most rudimentary techniques, which simply considers the existence of words (or *n*-grams) in a sentence, regardless of ordering. These create a sparse matrix. Term frequency-inverse document frequency (TF-IDF) is another popular vectorization technique that represents the uniqueness of an *n*-gram. Again, creating a sparse matrix.

The more recent popularization of word and sentence embeddings makes it possible to consider the context within which natural language is used. Word embeddings

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Table 4.1 Common Terms Used in Industrial-Organizational Psychology

Term	Acronym	Definition
Artificial Intelligence	AI	Methods that simulate intelligent human behavior, such as learning, decision making, sensing, and predicting.
Machine Learning	ML	A primary AI method including algorithms that detect patterns in data.
Supervised Machine Learning		An ML model trained to detect the relationships between variables and a criterion (ground truth), and then replicate those relationships on new, unseen data.
Unsupervised Machine Learning		An ML model that inductively uncovers patterns in the data; a human makes sense of the patterns; there is no criterion in unsupervised ML
Deep Learning		A subset of ML; uses neural networks and transformers that have many internal layers of analysis.
Transformers		A type of deep-learning architecture that advanced attentional mechanisms to improve understanding of human language, even if words are far apart in a sentence; more efficient than previous architectures (e.g., Recurrent Neural Networks).
Bag-of-Words	BoW	An NLP vectorization method that represents text data as unordered words (e.g., "woman leads people" is the same as "leads woman people"). It creates sparse matrices.
Word Embeddings		An NLP vectorization method that represents text data while considering syntax, unlike BoW. It creates more dense matrices.
Large Language Models	LLMs	Language-based models that generate text data; trained in millions or billions of parameters.
Natural Language Processing	NLP	Algorithms specifically designed to enable machines to understand and generate human language.
Black Box		Colloquialism to describe how the mathematical operations of some ML algorithms cannot be examined directly.
Ground Truth		Criterion used to train machine learning models to predict, such as past hiring decisions.
Features		Variables extracted and scored by a machine learning model; in NLP, this includes *n*-grams.
n-grams		Terms extracted from a corpus of text; can be single terms ("leader"), or multi-terms (e.g., bi-grams, "strong leader")

consider context insofar as they are able to help reduce issues of polysemy—where two words are spelled the same but mean different things. For example, the word "light" can refer to information ("brought to light"), tone of language ("we kept the conversation light"), or weight ("the box was light"). Otherwise, their extract of context is minimal. Sentence embeddings are much more useful in leveraging context because the level of analysis is the sentence rather than a single—or small set—of

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

terms. In addition to offering a more precise representation of language, embeddings can improve prediction by forming more dense matrices, rather than the sparse matrices produced by BoW. Sparse matrices are often quite large with many dimensions that are mostly zeros, which require more data to run, more power to detect reliable effects, and may be more computationally demanding, although they may also be more explainable (e.g., Drikvandi & Lawal, 2020; Hu et al., 2019; Krishnan et al., 2018; this has been referred to as the "curse of dimensionality," Bellman, 1957). Dense matrices, on the other hand, have fewer dimensions, are mostly nonzeros, and are therefore likely more efficient (i.e., less computationally demanding) (Hu et al., 2019). Embeddings create dense matrices by representing *n*-grams closer together that are more similar. Embeddings are often developed through transformers that learn the semantic similarities of *n*-grams (Mikolov et al., 2013; Reimers & Gurevych, 2020). This is analogous to the challenge we face when we dichotomize variables or have variables that are already dichotomized such that we can dramatically reduce variance making it difficult to detect stable patterns (Cohen, 1983). All the journal articles in this review used NLP.

Broadly, ML refers to algorithms that detect patterns in data. In supervised machine learning, algorithms not only detect those patterns, but replicate the patterns in unseen data. These models are built based on ground truths, which are typically continuous (e.g., human ratings) or binary (e.g., decisions, events). For example, Sajjadiani et al. (2019) built a model from text-based applicant data about their work history to predict their performance and likelihood of turnover. In this sample of teachers, the ground truths were a robust set of performance evaluations (other-rated evaluations, observations, and student performance on standardized tests) and turnover hazard ("whether and when the employee turned over; 2019, p. 1214). Meanwhile, Fan et al. (2023) built a chatbot to score applicant's personality from a semi-structured typed conversation with the chatbot. The ground truth they used to train the model was self-assessments of personality. It is notable to observe that our common analytic techniques of OLS and logistic regression can both be considered ML algorithms. Yet, this does not mean that all studies that use OLS regression are applying AI. Regression-based methods are common algorithms, but what makes a regression-based study AI is in its application. Simply running an OLS regression to uncover the coefficient representing the relationship among two variables while controlling others is not AI. Rather, applying those weights to new data to estimate a score would be considered AI. As an example from I-O psychology, Campion et al. (2016) used OLS regression to build a model on more than 40,000 job applications with text-based and numeric data. This model was trained on a continuous score, and those weights were applied in subsequent hiring waves to generate a machine-derived application score to inform hiring decisions. That this model was used on data it was not trained on makes it an application of AI.

In instances where there is no ground truth (e.g., human application score, hiring decision), researchers can use unsupervised machine learning, which is where the algorithm surfaces a pattern in the data and the human makes sense of that pattern (e.g., topic modeling). This method is analogous to factor or cluster analyses

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

on already quantified data (e.g., a newly developed scale). Although factor analyses are common in I-O psychology research, unsupervised machine learning as it pertains to text-based data is less common in what has been published in I-O psychology. However, it is becoming increasingly popular as it is recognized as a way to efficiently summarize text-based data, such as open-ended responses in surveys on candidate reactions and employee engagement. The two papers that uncovered employee reactions to work-from-home orders in 2020 are exemplars of this method (Min et al., 2021; Zhang et al., 2021). In another example, Al-Qallawi and Raghavan (2022) used unsupervised machine learning to understand how candidates reacted to game-based assessments, and Green et al. (2023) analyzed leadership narratives to generate a taxonomy of structural leadership cues.

Because AI is about simulating human intelligence, it may be easiest to think of ML as intending to imitate how the human brain processes information, particularly drawing from its hierarchical architecture (McCulloch & Pitts, 1943). Deep learning was developed to move us closer to this goal by introducing multiple layers through which information can be passed through and processed. This is inspired by how signals are transmitted between neurons through synapses in the human brain, and by the structure of these neural connections, which are referred to as artificial neural networks in AI (Huang, 2017). The strength of the connection between "neurons" in an AI model is represented by a parameter or weight (Agoston, 2024). Information is transmitted via weights within or among layers and can be fed forward (i.e., "feedforward processing") from input to output, or recursively where the outputs of a "neuron" become the input either to that same "neuron" or another "neuron" (i.e., recurrent neural networks) (Sazli, 2006). The model "learns" by comparing its output to the ground truth and generating a loss function representative of how much it must "learn" to achieve the ground truth (Elharrouss et al., 2025). One of the most significant advancements in deep learning models was the development and introduction of the BERT architecture: Bidirectional Encoders Representations from Transformers (Devlin et al., 2019). What made this architecture so significant was its leveraging of deep learning that improved performance, and its generalizability and accessibility to researchers more broadly to fine-tune it for their needs (Gardazi et al., 2025). Researchers have advanced variants of BERT including RoBERTa, or Robustly Optimized BERT Pretraining Approach. RoBERTa improved on BERT by training the model longer and with more data, removing tasks that did not help performance, and alternating what terms it hid during training to improve performance (Liu et al., 2019). Several papers in our review used deep learning. For example, Thompson et al. (2023) sought to train a model to score open-ended assessments based on subject matter expert (SME) ratings on seven competencies across three companies. They compared several algorithmic approaches (BoW, Long Short-Term Memory [LSTM], and RoBERTa) and found that RoBERTa outperformed the other methods. The performance of RoBERTa as a pre-trained model was enabled by this team's further training of the model, illustrating one of the benefits of deep learning models. Others in our review similarly leverage deep learning (e.g., Hardy et al.'s study in Koenig et al., 2023).

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Artificial Intelligence in Industrial-Organizational Psychology

To date, selection is the most popular application of AI in I-O psychology with 19 of the 36 papers explicitly on selection (see Table 4.2). Although, as illustrated by the subsequent chapters in this book, more I-O subdomains are using these advanced methods. For example, five papers have been published in the performance management domain. Together, they found that we can use NLP to score

Table 4.2 Overview of Published Research Using Artificial Intelligence in Industrial-Organizational Psychology

Article No.	Authors	Brief Description	I-O subdomains	Results
1	Campion et al. (2016), *JAP*	Trained a model to score accomplishment records from job applicants	Selection	• ML model can score accomplishment records as well as a human rater • Scores likely do not include larger subgroup differences than the data on which the model was trained
2	Speer et al. (2018), *PPsych*	Scored sentiment of performance narratives	Performance Management	• Narrative scores were reliable and correlated with numeric ratings • Narrative scores correlated with job performance
3	Banks et al. (2019), *JOM*	Analyzed recruiting signals from company websites	Recruitment	• Can uncover meaningful recruitment signals from company websites • There are differences in signal strength depending on whether the organization has domestic or international operations
4	Sajjadiani et al. (2019), *JAP*	Trained a model to score work histories and predict performance and turnover	Selection & Turnover	• Used machine learning to uncover themes in reasons for turning over from work history • Can score work histories to predict performance and turnover
5	Speer et al. (2019), *JBP*	Extended the Great Eight Competencies word dictionary	Performance Management	• Showed that researchers can develop and content validate dictionaries of job-related constructs (e.g., leading and deciding, adapting and coping) to score performance narratives

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Article No.	Authors	Brief Description	I-O subdomains	Results
6	Campion & Campion (2020), *RPHRM*	Reviewed the literature on uses of machine learning in I-O psychology	General I-O	• Showed the utility and functionality of AI across subdomains of I-O psychology
7	Speer (2021), *ORM*	Scored narratives for performance dimensions	Performance Management	• Scored content and valence of performance narratives • Scores converged with human ratings • Demonstrated generalizability of scoring method; generalizability was stronger on data with similar survey design
8	Al-Qallawi & Raghavan (2022), *IJSA*	Used NLP to uncover applicant reactions to game-based assessments (GBAs)	Selection	• Candidate view GBAs favorably • GBAs can be perceived as having less face validity
9	Guo et al. (2024), *HRMJ*	Offers two demonstrations of NLP methods	Selection	• Shows that BoW can be used to predict numerical ratings • Shows how a neural network scoring method can be used to predict self-report personality from situational interview questions
10	Kotlyar & Krasman (2022), *IJSA*	Scored teamwork skills from chatbot simulation	Performance Management	• ML-based simulation scores predicted peer ratings beyond self-report personality and teamwork situational judgment items • Women outscored men
11	Min et al. (2021), *JAP*	Used open-source data to evaluate responses to work-from-home orders during COVID-19	Employee Relations	• Showed how to score collective sentiment using deep learning
12	Zhang et al. (2021), *JAP*	Used open-source data to analyze sentiment of enforced work-from-home during COVID-19	Employee Relations	• Uncovered themes and sentiment of reactions to work-from-home orders

continued

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Table 4.2 *continued*

Article No.	Authors	Brief Description	I-O subdomains	Results
13	Hickman et al. (2022), *JAP*	Measured personality from automated video interviews (AVI)	Selection	• Showed how to automate scoring of video interviews • Tested differences in validity such that AVI personality assessments trained on interviewer ratings had higher validity than when trained on self-report data • Demonstrated that AVI personality assessments predicted academic outcomes
14	Kumar & Burns (2024), *PPsych*	Developed a model to predict occupational safety rates	Occupational Health & Safety	• Uncovered understudied predictors of workplace safety accidents
15	Campion & Campion (2023), *PPsych*	Synthesizes emerging best practices on using ML in selection	Selection	• Provides a tutorial on ML concepts • Identifies lessons learned from the special issue • Offers a list of emerging best practices on using ML in selection
16	Hernandez & Nie (2023)	Used GPT-2 to develop personality assessment items	Selection	• Showed how to use generative AI to create a valid assessment of personality • Found reliability and validity evidence equivalent to scales developed by traditional methods
17	Landers et al. (2023), *PPsych*	Conducted a simulation of machine learning models to uncover psychometric properties beyond OLS	Selection	• Found that machine learning models perform better with less data and a smaller sample-to-scale count ratio than OLS
18	Koenig et al. (2023), *PPsych*	*6 Studies* (Koenig et al.) Evaluated whether models developed to score responses from one set of prompts can be used to score responses from other prompts	Selection	• Found that algorithms can generalize from one prompt to another. • Optimal generalization occurs when the prompts are the same and content is similar.

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Article No.	Authors	Brief Description	I-O subdomains	Results
		(Yankov & Speer) Evaluated use of NLP to score text data from assessment centers	Selection	• Found that the model was as reliable as a human rater • The model had criterion-related validity similar to human raters or better
		(Hardy et al.) Tested the performance of deep learning models to score applicant responses	Selection	• Can use deep learning to score competencies in applicant responses to assessments • Deep learning score shared small relationships with supervisor ratings of job performance • Subgroup differences from deep learning scoring was in line with the literature
		(Liu et al.) Used machine learning to score audio responses from a simulation	Selection	• Can score transcribed audio responses with levels of reliability and criterion-related validity similar to a human rater • Scores predicted beyond existing assessments
		(Sun et al.) Evaluated the use of machine learning to predict multiple outcomes simultaneously	Selection	• Showed how to use machine learning to improve prediction on multiple outcomes (e.g., productivity and turnover) • Showed that model performance beyond traditional methods may only be modest when data are from highly structured methods
		(Lebanoff et al.) Evaluated the use of machine learning models to identify KSAs from job descriptions	Job Analysis	• Found that a model could be trained to extract KSAs as well as humans
19	Zhang et al. (2023), *PPsych*	*3 Studies* (Zhang et al.) Tested whether we can eliminate subgroup differences using machine learning	Selection	• Found that when eliminating subgroup differences in machine learning models, it reduces validity and can penalize high-performing racial minority candidates

continued

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Table 4.2 *continued*

Article No.	Authors	Brief Description	I-O subdomains	Results
		(Hickman et al.) Evaluated whether oversampling can help reduce subgroup differences in machine learning application score	Selection	• Found that removing subgroup differences in training data via oversampling increased adverse impact ratios and reduced validity, though effects were small
		(Song et al.) Developed an open-source tool for multi-objective optimization	Selection	• Showed how to use Pareto optimization to optimize model prediction of up to three objectives
20	Chowdhury et al. (2023), *HRMR*	Reviewed literature on organizational resources required to develop AI capability	General I-O	• Developed a framework of AI capabilities within an organization
21	Fan et al. (2023), *JAP*	Created a chatbot to measure applicant personality interactively	Selection	• Showed how to score personality from interactions with a chatbot • Chatbot-inferred personality scores showed a relationship with GPA beyond ACT scores
22	Green et al. (2023), *JAP*	Created a taxonomy of structural leadership cues	Leadership	• Uncovered leadership situations from leader narratives
23	Hickman et al. (2023), *IJSA*	Developed NLP-based models to score assessment center data	Selection	• Tested differences in convergent validity of different operationalization of verbal behavior • Showed how replacing a human assessor with an ML score did not meaningfully change criterion-related validity
24	Lee et al. (2023), *JBP*	Used GPT-3 to generate personality items	Selection	• Can use generative AI to create a valid assessment of personality
25	Putka et al. (2023), *JBP*	Used machine learning to automate job analysis	Job Analysis	• Can use NLP to identify important KSAs from job descriptions

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Article No.	Authors	Brief Description	I-O subdomains	Results
26	Rottman et al. (2023), *JAP*	Used machine learning to score responses to automated video interviews using multi-penalty optimization	Selection	• Tested different algorithms for performance on diversity-validity tradeoff and found that removing predictors that showed subgroup differences reduced subgroup differences, but also validity
27	Thompson et al. (2023), *JBP*	Used deep learning to score competencies from virtual assessment center exercises	Selection	• Can score competencies of applicants using deep learning • Achieved validity nearly identical to human raters • Some differences in gender and racial subgroups by algorithm
28	Wang et al. (2023), *JBP*	Applied topic modeling to uncover themes around DEI policies in organizations	Recruitment	• Employees rated organizations more highly on diversity and inclusion if they communicate identity-consciousness
29	Speer et al. (2024), *JAP*	Created set of dictionaries to score worker attitudes (Text-Based Attitude and Perception Scoring [TAPS] dictionaries)	Performance Management	• Develop language models to measure 28 work attitudes • Can measure attitudes using valence (sentiment) and theme (content) scores • Attitudes correlated with self-report measures of organizational citizenship behaviors and counterproductive work behaviors • Prediction is stronger when prompts were more specific to the construct of interest
30	Song et al. (2024), *PPsych*	Used machine learning to predict career choice	Career Development	• Showed how a model could be built to predict occupational membership and vocational aspirations based on interest inventories

continued

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Table 4.2 *continued*

Article No.	Authors	Brief Description	I-O subdomains	Results
31	Campion et al. (2024), *JAP*	Trained a machine learning model to score text fields in applications	Selection	• Showed that the model could score open-ended responses as well as a human rater. • Demonstrated how the inclusion of text-based data reduced subgroup differences in total application score • Predicted selection rating beyond mental ability tests and performance beyond selection rating
32	Chekili & Hernandez (2024), *ORM*	Built a model to classify participant ethnicity and gender	Measurement	• Demonstrated how to build a model to classify a person's gender and ethnicity from open-source data (e.g., websites, mobile apps)
33	Hickman, et al. (2024), *IJSA*	Tested how LLMs performed on quantitative and verbal ability assessments	Selection	• LLMs performed well with verbal ability, but very poorly for quantitative ability • Prompt and temperature parameters showed little differences in scores
34	Liff et al. (2024), *JAP*	Used machine learning to score competencies of responses to AVIs	Selection	• Can score competencies of applicants using machine learning • Machine learning scores converged with human ratings
35	Min et al. (2024), *PPsych*	Built a model to predict turnover and worker destination	Turnover	• Demonstrated ability to train model to predict turnover • Recommender system also predicted where the worker would end up
36	Stevenor et al. (2024), *IJSA*	Trained a machine learning model to infer applicant personality	Selection	• Trained a machine learning model on data from low-stakes contexts and applied the model to high-stakes interviews • Model enabled inference of personality • Scores had weak-moderate relationships with supervisor ratings of job performance

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Article No.	Authors	Brief Description	I-O subdomains	Results
37	Song et al. (2024), *JAP*	Analyzed the nature of the relationship between personality (Five Factor Model) and job performance	Measurement	• Used "interpretable AI" to analyze the relationship between personality and job performance • When modeled as facets, the non-linear model was more predictive of performance than the linear model
38	Hickman et al. (2025), *JAP*	Systematically reviewed bias in machine learning	Measurement	• Identified sources of algorithmic bias and where they present in a four-stage framework of ML model production • Describes opportunities to reduce bias at each stage

Note: Journal acronyms are as follows: *HRM = Human Resource Management Journal; HRMR = Human Resource Management Review; ISJA = International Journal of Selection and Assessment; JAP = Journal of Applied Psychology; JBP = Journal of Business and Psychology; JOM = Journal of Management; ORM = Organizational Research Methods; PPsych = Personnel Psychology; RPHRM = Research in Personnel and Human Resources Management*

applicant narratives to supplement performance scores, develop generalizable scoring methods through inductive approaches (Speer, 2021) and deductive approaches (e.g., word dictionaries; Speer et al., 2019; Speer et al., 2024), and predict peer teamwork ratings (Kotlyar & Krasman, 2022). While all the papers on AI focus on measurement, only three papers in our review are categorized as such because their focus is explicitly on measurement and their results are applicable to a number of I-O subdomains. For example, Song et al. (2024) used AI to dig deeper into the nature of the relationship between the Five Factor Model of personality and performance and found that the best performing model was non-linear at the facet level. Hickman et al. (2025) reviewed the literature on algorithmic bias and identified where researchers can take steps to mitigate bias at each of the four stages of their framework.

Two papers have used AI to improve our understanding of employee engagement and sentiment. Min et al. (2021) and Zhang et al. (2021) uncovered employee attitudes about work-from-home mandates from open data sources (e.g., X, formerly known as Twitter) and found evidence of both positive and negative reactions to these orders. Two papers have shown how AI can be used to inform recruitment. In the first, Banks et al. (2019) uncovered recruitment signals from company websites. In the second, Wang et al. (2023) evaluated individuals' responses to organizational diversity, equity, and inclusion (DEI) statements from Glassdoor. While not explicitly on recruitment, the method of using open-source data such as Glassdoor

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

to understand the perceptions of one's talent pool holds promise for organizations. Finally, two papers have been built to predict turnover (Min et al., 2024; Sajjadiani et al., 2019), two papers illustrated how to use AI in job analysis (Putka et al., 2023; Lebanoff et al.'s study in Koenig et al., 2023), one paper built a model to predict occupational accident rates (Kumar & Burns, 2024), and one paper built a model to predict career choice (Song et al., 2024).

An important contribution of this family of methods to date has been how NLP has enabled I-O psychologists to systematically score candidate text-based data. Structured, reliable, efficient, and fair processes for scoring applicant text data has long-plagued selection scholars and practitioners (Campion & Campion, 2023). Candidates produce text data in various forms—resumes, cover letters, application blanks, essays, and so on. Scoring these data requires training and frame-of-reference meetings to ensure reliability among raters; it requires structured rubrics to support job-related evaluations and reduce non-job-related rater preferences; and it requires a significant amount of time for human raters to score all applicants. Moreover, data are often unstandardized such that resumes and the like may not provide exactly the same information in the same format, which may influence human evaluation (e.g., Schramm & Dortch, 1991). A well-validated model solves these challenges and offers additional benefits such as providing candidates opportunities to discuss their knowledge, skills, and abilities (KSAs) in greater detail, and text-based data may be the optimal way to capture constructs that have been traditionally difficult to scores using already quantified measures (e.g., motivation and work experience; Campion et al., 2024; Tesluk & Jacobs, 1998; Van Iddekinge et al., 2019).

The increased focus on natural language has created two cascading contributions: the adverse impact-validity tradeoff and LLMs. The adverse impact–validity tradeoff refers to the historical finding that the most predictive methods tend to yield larger subgroup differences that often result in adverse impact (Ployhart & Holtz, 2008). At least one paper has demonstrated that the use text-based data empowers applicants to provide detailed descriptions of experiences through which they have cultivated job-related KSAs (Campion et al., 2024). The representation of these KSAs in text is important because this information is not otherwise documented or evaluated in traditional materials, such as resumes or captured in historic assessments. As such, these researchers found that inclusion of this information not only improved prediction of human ratings and job performance, but also reduced subgroup differences in the total application score (Campion et al., 2024). Application blanks and other opportunities for applicants to more broadly discuss the life experiences in which they developed job-related skills are promising methods for addressing this enduring challenge in selection.

The second contribution is an emerging one with significantly more scholarship needed. That is the use of LLMs. For most researchers, LLMs still feel like the "Wild West." However, nascent scholarship in I-O psychology suggests there is promise. Hernandez and Nie (2023) and Lee et al. (2023) used GPT-2 and GPT-3,

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

respectively, to generate items to capture personality. Both evaluated the reliability, structure, and validity of these items. These results collectively support the use of such methods to generate items used in various I-O domains (e.g., selection, performance management, career development). Further, they illustrate that what is generated is amenable to traditional scale evaluation best practices from I-O psychology by evaluating the psychometrics of what LLMs create. Item writing is a time-consuming task and is a logical example of where I-O researchers and practitioners can begin to save labor by using ML. LLMs have also increased the accessibility of AI to non-experts because they require no programming skills.

A decade ago, these methods seemed to be inaccessible to individuals not formally trained in computer science, data science, data engineering, and related disciplines. "Coding" in I-O psychology generally referred to manual content analyses, manually extracting information from manuscripts to conduct meta-analyses or narrative reviews, or deconstructing text-based data for grounded theory; whereas "writing syntax" meant you were writing command prompts in SPSS, SAS, Stata, or even R (which was on the rise in our science at the time and is free) to run analyses. With the continued popularity of Python as the primary AI language, "coding" now refers to all these activities plus writing Python code to run ML models. As a common language in the home disciplines of AI with free resources enabled by crowdsourcing via GitHub and other repositories, Python has provided access for researchers in I-O psychology and related disciplines to leverage these advanced methods. Although it is not uncommon for researchers to use both Python and R in the same research (e.g., Fan et al., 2023), the advent and popularization of LLMs also represents a form of increased access. These tools can be used to improve efficiencies in daily tasks (e.g., writing job descriptions, generating interview questions, writing assistance), and they have also become an invaluable resource to help researchers learn to code. Nevertheless, new issues are emerging with the use of LLMs, such as whether and how candidates use LLMs, and whether LLMs create fertile ground for organizational over-reliance on these tools, which we discuss below.

Emergent Challenges

At present, two of the largest challenges I-O researchers face when using AI is over-reliance on these methods and explainability. Broadly, *over-reliance on AI* refers to excessive dependence on these tools for the purposes of making decisions and completing tasks. Under the name of "automation bias," researchers have been tuned into the issues of using automated systems for decades citing disastrous or near-disastrous examples of humans over-relying on technology particularly in aviation (Parasuraman et al., 1993; Skitka et al., 1999). Automation bias is defined generally as a human's tendency to trust the AI tool over their own experience or instinct (e.g., Cheng, 2024). The core mechanism driving this over-reliance and bias is the

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

"allocation of limited user attention" (Lyell & Coiera, 2017, p. 423), thus reducing the likelihood of the human user identifying problems in automated systems. It is expected that humans will find ways to manage the immense amount of cognitive stimulation that work demands by seeking ways to quickly make decisions (Fiske & Taylor, 1994). This idea is central to perceptual/motor job design, which Campion (1988, p. 468) describes as:

> oriented toward human mental capabilities and limitations, primarily with regard to attention and concentration requirements of jobs. This approach relates favorably to reliability outcomes (e.g., reduced error- and accident-likelihoods) and positive user reactions (e.g., reduced mental overload, fatigue, stress, and boredom as well as favorable attitudes toward work stations and equipment).

Although reducing cognitive strain on workers may create a more favorable work environment as workers need to exert less energy into their labor, the integration of automation has expectedly created a risk of over-reliance. The most concerning proximal outcome of over-reliance in practice is error: erroneously flying a plane, erroneously missing a diagnosis, and in I-O psychology, erroneously allowing the AI tool to make high-stakes decisions (e.g., hiring, promoting, firing). This error can have immediate implications of relying on inaccurate information. Perhaps more alarming, however, is the charge that over-reliance will reduce human's impulse to think critically. This concern has been voiced most prominently when predicting the impact of AI on education. For example, in legal education, Peoples (2025, p. 1) concluded that there are "pitfalls of over-reliance on artificial intelligence (AI) resulting in a loss of logic, reasoning, and critical thinking skills." In pathology, Cheng (2024, p. 8) wrote, "Some groups have voiced concerns that over-reliance on AI tools for pathology case sign-out may lead to deskilling, burnout, and diminished knowledge of the histology and mechanisms of disease."

I-O psychologists must face the same concern of over-reliance for at least three reasons. First, it is our duty to measure and evaluate the KSAs of candidates to select those that can perform the job. As such, over-reliance that may result in reduced critical thinking required of highly demanding jobs may have implications for how we assess candidates. Second, we have yet to understand the implications of using models to score candidate data wherein the candidate data were created by the same model. Put differently, if a candidate uses ChatGPT to create their resume, and that resume is scored using GPT-powered tools, does this undermine the validity of the score? This issue relates to a broader concern about "model collapse," which refers to "a degenerative process affecting generations of learned generative models, in which the data they generate end up polluting the training set of the next generation. Being trained on polluted data, they then mis-perceive reality" (2024, p. 755). This means that LLMs continue to be trained on open-sourced online data, which will increasingly be written by LLMs. Shumailov et al. (2024, p. 755) claim this "causes

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

irreversible defects . . . in which the tails of the original content distribution disappear." As we continue to deepen our understanding of how we can use LLMs to score text-based candidate data, it is imperative we simultaneously pursue interventions to reduce model collapse.

Third, we need to continually reinforce the shared perspective among I-O psychologists that AI-based tools should be used to inform human decision making rather than wholly replace it (e.g., Campion & Campion, 2024). Efforts to do this may include using these tools to create tiers of applicants or employees according to their score and ensuring final decisions are made by humans such as hiring, pay increases, promotions, and the like (e.g., Campion et al., 2016; Campion et al., 2024); ensuring face and content validity when evaluating items developed by generative AI tools (e.g., Hernandez & Nie, 2023; Lee et al., 2023); and continually monitoring model generalizability to ensure models developed in one context are defensible in another (e.g., Koenig et al., 2023; Speer et al., 2024). We are only capable of defending our AI-based methods if we understand them, which requires us to eschew our own impulse to over-rely on these tools and be able to *explain* them.

While explainability will be covered in a dedicated chapter, we surface the notion because it is relevant to the continued use of LLMs and is an antidote to over-reliance. Explainable AI (XAI), is defined as "a collection of approaches that seek to make AI systems more interpretable to stakeholders in ways that satisfy their needs to understand the functioning and outputs of AI systems" (Pitcher & Wang, this book; see also Langer, Oster, et al., 2021). In many ways, this is no different than explaining any model produced in I-O psychology: describe the data quality, the model building steps and decisions, the size and quality of the relationships among predictors and outcomes, and if applicable, how a particular set of materials (e.g., application) was scored (Hickman et al., 2025; Song et al., 2024). However, because these methods are interpreted as more opaque than our traditional statistical approaches (e.g., bivariate correlations), they require additional evidence to ensure the results are trustworthy, valid, and defensible.

Conclusion

ML is an unprecedented opportunity to improve I-O psychology and it will have profound effects. However, it must come with understanding and not blind us without understanding the technology, how to use it correctly, and how to explain it to the various stakeholders. Our review suggests that I-O psychologists have already begun to embrace these methods. Cleverly, they have found ways to use AI to address resource constraints that have prohibited organizations' ability to systematically, efficiently, and fairly score text-based application data to improve selection decisions. I-O psychologists must continue to adopt these methods or other fields will dominate our science and practice without the necessary domain knowledge.

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

References

Agoston, D. V. (2024). Of artificial intelligence, machine learning, and the human brain. Celebrating Miklós Palkovits' 90th birthday. *Frontiers in Neuroanatomy*, *18*, article 1374864. https://doi.org/10.3389/fnana.2024.1374864

Al-Qallawi, S., & Raghavan, M. (2022). A review of online reactions to game-based assessment mobile applications. *International Journal of Selection and Assessment*, *30*(1), 14–26.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., Frear, K. A., Berka, G., Heggestad, E. D., & Gordon, H. L. (2019). Strategic recruitment across borders: An investigation of multinational enterprises. *Journal of Management*, *45*(2), 476–509.

Bellman, R. E., & Rand Corporation (1957). *Dynamic programming*. Princeton University Press.

Campion, E. D., & Campion, M. A. (2020). Using computer-assisted text analysis (CATA) to inform employment decisions: Approaches, software, and findings. In M. R. Buckley, A. R. Wheeler, J. E. Baur, & J. R. B. Halbesleben (Eds.), *Research in personnel and human resources management* (pp. 285–325). Emerald Insight.

Campion, E. D., & Campion, M. A. (2024). Impact of machine learning on personnel selection. *Organizational Dynamics*, *53*(1), 101035.

Campion, E. D., Campion, M. A., Johnson, J., Carretta, T. R., Romay, S., Dirr, B., Deregla, A., & Mouton, A. (2024). Using natural language processing to increase prediction and reduce subgroup differences in personnel selection decisions. *Journal of Applied Psychology*, *109*(3), 307–338.

Campion, M. A. (1988). Interdisciplinary approaches to job design: A constructive replication with extensions. *Journal of Applied Psychology*, *73*(3), 467–481.

Campion, M. A., & Campion, E. D. (2023). Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research. *Personnel Psychology*, *76*(4), 993–1009.

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology*, *101*(7), 958–975.

Chekili, A., & Hernandez, I. (2024). Demographic inference in the digital age: Using neural networks to assess gender and ethnicity at scale. *Organizational Research Methods*, *27*(2), 301–328.

Cheng, J. (2024). Applications of large language models in pathology. *Bioengineering*, *11*(4), 342.

Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., & Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review*, *33*(1), 100899.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(3), 249–253.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference*

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

of the North American chapter of the association for computational linguistics: Human language technologies (NAACL-HLT) Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1810.04805

Drikvandi, R., & Lawal, O. (2020). Sparse principal component analysis for natural language processing. *Annals of Data Science*, *10*(1), 25–41. https://doi.org/10.1007/s40745-020-00277-x

Elharrouss, O., Mahmood, Y., Bechqito, Y., Serhani, M.A., Badidi, E., Riffi, J., & Tairi, H. (2025). Loss functions in deep learning: A comprehensive review. *arXiv preprint*. arXiv:2504.04242

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. (2023). How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. *Journal of Applied Psychology*, *108*(8), 1277–1299.

Fiske, S. T., & Taylor, S. E. (1994). *Social Cognition* (2nd ed.). McGraw-Hill.

Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., & Alshemaimri, B. (2025). BERT applications in natural language processing: A review. *Artificial Intelligence Review*, *58*(6), article 166. https://doi.org/10.1007/s10462-025-11162-5

Green, J. P., Dalal, R. S., Fyffe, S., Zaccaro, S. J., Putka, D. J., & Wallace, D. M. (2023). An empirical taxonomy of leadership situations: Development, validation, and implications for the science and practice of leadership. *Journal of Applied Psychology*, *108*(9), 1515.

Guo, F., Gallagher, C. M., Sun, T., Tavoosi, S., & Min, H. (2024). Smarter people analytics with organizational text data: Demonstrations using classic and advanced NLP models. *Human Resource Management Journal*, *34*(1), 39–54.

Hernandez, I., & Nie, W. (2023). The AI-IP: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, *76*(4), 1011–1035.

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, *107*(8), 1323.

Hickman, L., Dunlop, P. D., & Wolf, J. L. (2024). The performance of large language models on quantitative and verbal ability tests: Initial evidence and implications for unproctored high-stakes testing. *International Journal of Selection and Assessment*, *32*(4), 499–511. https://doi.org/10.1111/ijsa.12479

Hickman, L., Herde, C. N., Lievens, F., & Tay, L. (2023). Automatic scoring of speeded interpersonal assessment center exercises via machine learning: Initial psychometric evidence and practical guidelines. *International Journal of Selection and Assessment*, *31*(2), 225–239.

Hickman, L., Huynh, C., Gass, J., Booth, B., Kuruzovich, J., & Tay, L. (2025). Whither bias goes, I will go: An integrative, systematic review of algorithmic bias mitigation. *Journal of Applied Psychology*, *110*(7), 979–1000. https://doi.org/10.1037/apl0001255

Hiebl, M. R. (2023). Sample selection in systematic literature reviews of management research. *Organizational Research Methods*, *26*(2), 229–261.

Hu, W., Wang, M., Liu, B., Ji, F., Chen, H., Zhao, D., Ma, J., & Yan, R. (2019). Transformation of dense and sparse text representations. *arXiv preprint*. arXiv:1911.02914.

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Huang, T.-J. (2017). Imitating the brain with neurocomputer: A "new" way towards artificial general intelligence. *International Journal of Automation and Computing*, *14*(5), 520–531. https://doi.org/10.1007/s11633-017-1082-y

Koenig, N., Tonidandel, S., Thompson, I., Albritton, B., Koohifar, F., Yankov, G., Speer, A., Hardy, J. H. III, Gibson, C., Frost, C., Liu, M., McNeney, D., Capman, J., Lowery, S., Kitching, M., Nimbkar, A., Boyce, A., Sun, T., Guo, F., . . . Newton, C. (2023). Improving measurement and prediction in personnel selection through the application of machine learning. *Personnel Psychology*, *76*(4), 1061–1123.

Kotlyar, I., & Krasman, J. (2022). Virtual simulation: New method for assessing teamwork skills. *International Journal of Selection and Assessment*, *30*(3), 344–360.

Krishnan, R., Liang, D., & Hoffman, M. (2018, March). On the challenges of learning with inference networks on sparse, high-dimensional data. In *International conference on artificial intelligence and statistics* (pp. 143–151). PMLR.

Kumar, L. S., & Burns, G. N. (2024). Determinants of safety outcomes in organizations: Exploring O∗ NET data to predict occupational accident rates. *Personnel Psychology*, *77*(1), 133–162. https://doi.org/10.1111/peps.12560

Landers, R. N., Auer, E. M., Dunk, L., Langer, M., & Tran, K. N. (2023). A simulation of the impacts of machine learning to combine psychometric employee selection system predictors on performance prediction, adverse impact, and number of dropped predictors. *Personnel Psychology*, *76*(4), 1037–1060.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473.

Lee, P., Fyffe, S., Son, M., Jia, Z., & Yao, Z. (2023). A paradigm shift from "human writing" to "machine generation" in personality test development: An application of state-of-the-art natural language processing. *Journal of Business and Psychology*, *38*(1), 163–190.

Liff, J., Mondragon, N., Gardner, C., Hartwell, C. J., & Bradshaw, A. (2024). Psychometric properties of automated video interview competency assessments. *Journal of Applied Psychology*, *109*(6), 921–948. https://doi.org/10.1037/apl0001173

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*. arXiv:1907.11692.

Lyell, D., & Coiera, E. (2017). Automation bias and verification complexity: A systematic review. *Journal of the American Medical Informatics Association*, *24*(2), 423–431.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4), 115–133.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*. arXiv:1301.3781.

Min, H., Peng, Y., Shoss, M., & Yang, B. (2021). Using machine learning to investigate the public's emotional responses to work from home during the COVID-19 pandemic. *Journal of Applied Psychology*, *106*(2), 214–229.

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Min, H., Yang, B., Allen, D. G., Grandey, A. A., & Liu, M. (2024). Wisdom from the crowd: Can recommender systems predict employee turnover and its destinations? *Personnel Psychology*, *77*(2), 475–496.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *International Journal of Aviation Psychology*, *3*(1), 1–23.

Peoples, L. (2025). Artificial intelligence and legal analysis: Implications for legal education and the profession. *arXiv preprint*. arXiv:2502.03487.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity–validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology*, *61*(1), 153–172.

Putka, D. J., Oswald, F. L., Landers, R. N., Beatty, A. S., McCloy, R. A., & Yu, M. C. (2023). Evaluating a natural language processing approach to estimating KSA and interest job analysis ratings. *Journal of Business and Psychology*, *38*(2), 385–410.

Reimers, N., & Gurevych, I. (2020). The curse of dense low-dimensional information retrieval for large index sizes. *arXiv preprint*. arXiv:2012.14210.

Rottman, C., Gardner, C., Liff, J., Mondragon, N., & Zuloaga, L. (2023). New strategies for addressing the diversity–validity dilemma with big data. *Journal of Applied Psychology*, *108*(9), 1425–1444.

Sajjadiani, S., Sojourner, A. J., Kammeyer-Mueller, J. D., & Mykerezi, E. (2019). Using machine learning to translate applicant work history into predictors of performance and turnover. *Journal of Applied Psychology*, *104*(10), 1207–1225.

SAS. (2019). Machine learning: What it is and why it matters. https://www.sas.com/en_us/insights/analytics/machine-learning.html

Sazlı, M. H. (2006). A brief review of feed-forward neural networks. *Communications Faculty of Sciences University of Ankara Series A2-A3*, *50*(1), 11–17.

Schmidgall, S., Ziaei, R., Achterberg, J., Kirsch, L., Hajiseyedrazi, S.P., & Eshraghian, J. (2024). Brain inspired learning in artificial neural networks: A review. *APL Machine Learning*, *2*(2), 021501. https://doi.org/10.1063/5.0186054

Schramm, R. M., & Neil Dortch, R. (1991). An analysis of effective resume content, format, and appearance based on college recruiter perceptions. *Bulletin of the Association for Business Communication*, *54*(3), 18–23.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, *631*(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human–Computer Studies*, *51*(5), 991–1006.

Song, Q. C., Shin, H. J., Tang, C., Hanna, A., & Behrend, T. (2024). Investigating machine learning's capacity to enhance the prediction of career choices. *Personnel Psychology*, *77*(2), 295–319.

Song, Q., Oh, I. S., Kim, Y., & So, C. (2024). Revisiting the nature and strength of the personality–job performance relations: New insights from interpretable machine learning. *Journal of Applied Psychology*, *110*(1), 1–26. https://doi.org/10.1037/apl0001218

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, *71*(3), 299–333.

Speer, A. B. (2021). Scoring dimension-level job performance from narrative comments: Validity and generalizability when using natural language processing. *Organizational Research Methods*, *24*(3), 572–594.

Speer, A. B., Perrotta, J., Tenbrink, A. P., Wegmeyer, L. J., Delacruz, A. Y., & Bowker, J. (2024). Turning words into numbers: Assessing work attitudes using natural language processing. *Journal of Applied Psychology*, *108*(6), 1027–1045.

Speer, A. B., Schwendeman, M. G., Reich, C. C., Tenbrink, A. P., & Siver, S. R. (2019). Investigating the construct validity of performance comments: Creation of the great eight narrative dictionary. *Journal of Business and Psychology*, *34*, 747–767.

Stevenor, B. A., Hickman, L., Zickar, M. J., Wimbush, F., & Beck, W. (2024). Validity evidence for personality scores from algorithms trained on low-stakes verbal data and applied to high-stakes interviews. *International Journal of Selection and Assessment*, *32*(4), 544–560.

Tesluk, P. E., & Jacobs, R. R. (1998). Toward an integrated model of work experience. *Personnel Psychology*, *51*(2), 321–355.

Thompson, I., Koenig, N., Mracek, D. L., & Tonidandel, S. (2023). Deep learning in employee selection: Evaluation of algorithms to automate the scoring of open-ended assessments. *Journal of Business and Psychology*, *38*(3), 509–527.

Tietjen, J. S. (2022). Amazing Grace—Computer pioneer Admiral Grace Murray Hopper. In A. E. Smith (Ed.), *Women in computational intelligence: Key advances and perspectives on emerging topics* (pp. 1–20). Springer International Publishing.

Van Iddekinge, C. H., Arnold, J. D., Frieder, R. E., & Roth, P. L. (2019). A meta-analysis of the criterion-related validity of prehire work experience. *Personnel Psychology*, *72*(4), 571–598.

Wang, W., Dinh, J. V., Jones, K. S., Upadhyay, S., & Yang, J. (2023). Corporate diversity statements and employees' online DEI ratings: An unsupervised machine-learning text-mining analysis. *Journal of Business and Psychology*, *38*(1), 45–61.

Zhang, C., Yu, M. C., & Marin, S. (2021). Exploring public sentiment on enforced remote work during COVID-19. *Journal of Applied Psychology*, *106*(6), 797–810.

Zhang, N., Wang, M., Xu, H., Koenig, N., Hickman, L., Kuruzovich, J., Ng, V., Arhin, K., Wilson, D., Song, Q. C., Tang, C., Alexander, L., & Kim, Y. (2023). Reducing subgroup differences in personnel selection through the application of machine learning. *Personnel Psychology*, *76*(4), 1125–1159.

Downloaded from academic.oup.com/book/62620/chapter/561204543 by OUP site access user on 12 October 2026

Emily D. Campion and Michael A. Campion, *Overview of Artificial Intelligence in Industrial-Organizational Psychology*. In: *Artificial Intelligence for I-O Psychologists*. Edited by: Isaac Thompson, Georgi P. Yankov, and Ivan Hernandez, Oxford University Press. © Oxford University Press (2026). DOI: 10.1093/9780197807309.003.0004